DOCUMENT RESUME

ED 256 499 PS 015 113

AUTHOR Fagot, Beverly I.; Hagan, Richard
TITLE Coding of Interactions: Is Reliability Really a Problem?
SPONS AGENCY National Inst. of Mental Health (DHEW), Bethesda, Md.
PUB DATE Apr 85
GRANT NIMH-5-R01-MH-37511-02
NOTE 26p.; Paper presented at the Biennial Meeting of the Society for Research in Child Development (Toronto, Ontario, Canada, April 25-28, 1985).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Automation; *Classroom Observation Techniques; *Data Collection; *Reliability; *Researchers; Research Methodology; *Research Problems; *Training; Young Children
IDENTIFIERS *Coding; Social Interaction

ABSTRACT

Covert checks of observational methodology reveal declines in reliability of observations. This appears to be particularly true when complex codes are used to track social interaction. The present study was undertaken to see whether reliability could be maintained through a combination of technological advancements and the development of improved observer training and maintenance procedures. Eight people were enrolled in a highly refined, 150-hour training program developed at the University of Oregon's Child Research Laboratory. Using automated data collectors, the trainees learned to code the interactions of young children. Following completion of the training program, observers remained in close contact with the project director through weekly meetings, allowing for discussion of unusual circumstances, performance feedback, and refinement of procedures. Subsequent covert checks failed to indicate any decline in agreement when scores were compared with data from overt checks. While agreement was high under all circumstances, this does not mean that complex coding systems should be applied in all observational studies. However, findings do suggest that when an appreciation of the subtleties inherent in social process are sought, complex systems are essential. (Author/RH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED256499

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
X This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

Coding of Interactions: Is Reliability Really a Problem?

Beverly I. Fagot and Richard Hagan

University of Oregon

and

Oregon Social Learning Center

This research was supported by a grant from NIMH (5 R01 MH 37511-02).

All correspondence should be sent to Beverly I. Fagot, Oregon Social Learning Center, 207 East 5th Avenue, Suite 202, Eugene, Oregon 97401.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Richard Hagan
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

PS 015113

Running Head: CODING OF INTERACTIONS

Abstract

Observational methodology has been recurringly troubled by declines in reliability whenever checks are made covertly. This appears to be particularly true when complex codes are used to track social interaction. This study was undertaken to see if reliability could be maintained through a combination of technological advancements and the development of improved observer training and maintenance procedures. Eight people were put through a highly refined, 150-hour training program developed at the University of Oregon's Child Research Laboratory. Using automated data collectors, the trainees learned to code the interactions of young children. Following completion of this program, observers remained in close contact with the project director through weekly meetings, allowing for discussion of unusual circumstances, performance feedback, and refinement of procedure. Subsequent covert checks of agreement failed to indicate any decline when compared with data from overt checks. Agreement was high under all circumstances. This does not, of course, mean that complex systems should be applied under all circumstances, but where an appreciation of the subtleties inherent in social process are sought, complex systems are seen as essential.

Coding of Interactions: Is Reliability Really a Problem?

Weinrott and Jones (1984), in a well-designed study of observer agreement using trained observers from four different projects, found a precipitous drop in agreement when assessments were made covertly. They also found that there was a relationship between the complexity of the coding system and the degree of drop in agreement when observers did not know they were being checked. The methodology of the study is outstanding, but we believe their conclusions concerning the direction that coding of interactions should take are now outdated. They concluded that, while comprehensive recording is clinically and theoretically appealing, it is of no benefit when it yields unreliable measures. The assumption seems to be that there is no way of correcting the drop in observer reliability under covert conditions.

We believe that the conclusion drawn by Weinrott and Jones is outmoded and no longer appropriate for the following reasons. First, the data in the study were collected several years ago, and the technology of the field has changed tremendously. Automated data collectors are now routinely available, and, perhaps even more important, small laboratory computers are commonplace, which provide researchers with immediate information on problems of reliability. Hence, feedback can be transmitted to observers while observations are still fresh in their minds.

Second, it is our feeling that we have learned a great deal about training observers over the past 10 years. Despite the fact that the observers in Weinrott and Jones' study were the some of the best trained in the country at the time, now one can do better.

Observer training is an ongoing process that is never ending. We are talking about a skill that calls for rote memory, rapid decision making, and fast motor reactions, and individuals who engage in such duties know that it is a never ending effort to keep such skills sharp. Observers should be regarded as very expensive, finely honed tools, and as such, they deserve as much care as the rest of our sensitive laboratory equipment.

We agree with Reid (1982) that one ought to use the simplest coding system possible for any particular question. On the other hand, one should not use a simple system just because reliability of a more complicated and interesting system is harder to maintain. Instead, one should put more time and work into training and skill maintenance to ensure data that are not only reliable but also best suited to the particular research questions.

We would like to present the training and maintenance procedures used in our project on the origins of mental health. A complicated code system is involved which allows us to look at a variety of situations. Yet we have been able to maintain a high degree of reliability among observers, measured both

covertly and overtly. We believe that information concerning interactive processes is worth the effort required to keep our observers working at peak performance.

Procedures for Observer Training and Maintenance

The Training Program

The procedures described in this section are an integral part of later reliability maintenance. As these procedures constitute a relatively complex and time consuming program, their practicality also deserves comment and will be discussed later in this paper. Initial training with our system requires approximately 150 hours of training. We have found that extended group training sessions of more than three hours are not advisable: trainee(s) and trainer alike become inefficient (use of time, performance) during longer sessions. However, additional individual practice sessions are helpful. Reliability is usually achieved in the seventh or eighth week of training.

Orientation. Observers are welcomed to the project with a general orientation program. An overview of the significance of observing and observational methods is given. They are provided with a set of research articles in an orientation packet that deals with various aspects of observing requirements. The packet also includes a copy of the interactive behavior code manual and the OS-3 microcomputer manual. Both documents are described later. As part of the orientation session, trainees cod

videotaped segments without the assistance of a systematic code. They are given immediate feedback on the strategies they appear to have used, along with a preset list of alternative strategies that might have been used. Trainees report this exercise helps give them a better and more immediate appreciation of the value and necessity of a systematic code, and counteracts their tendency to center on a set code's limitations rather than its advantages.

Coding description and procedures for mastery. The behavioral coding system used in our research provides descriptions of play preferences of young children, the individuals with whom they interact, and the type of reactions elicited, if any. The code is primarily used to give running accounts of a target child's behavior over a continuous period of time, although it has also been used in scan sampling studies (Fagot, 1978; Fagot & Hagan, 1984).

The code is broken down into five subsets, totaling eight digits. First, the context is coded with one of fourteen context codes (two digits, 01-14). It refers to the environment relevant to the target child's behavior; hence it helps give a picture of the general situation occurring. Examples of context codes are Fantasy Play, Musical Activity, Art Activities, and Building Toys. Second, the interactive nature of the child's behavior is coded with one of fourteen interactive codes (two digits,

01-14). This portion of the code is designed to describe whom, if anyone, the target child's behavior is directed toward. Examples of interactive codes are Babble, Use Language, and Helping Behaviors. Third, the recipient of the child's interactive behavior is coded (one digit, 1-8). The recipient is the person or persons that are the focus of the target child's behavior. Possible recipients include male or female children or adults, or any combination thereof. Fourth, the reactor is coded (one digit, 1-8). The reactor is that person or persons who respond to the behavior of the target child. Reactor categories are identical to the recipient categories. Fifth, the reaction of the reactor is recorded (two digits, 01-14). This portion of the code is designed to identify the type of response the reactor gives to the target child's behavior or any initiations directed toward the child. Examples of reaction codes are Comment Favorably, Physical Restraint, and Verbal Interaction.

As the code is applied continuously during an observational session, a change in any of the five portions of the code will result in a new entry being recorded by the observer.

Rapid memorization of code definitions and items is emphasized from the start. Within two days of the initial orientation, trainees start receiving quizzes on various code items and their meanings. Several methods are used to test both recall and recognition. First, trainees are tested on individual

code items. One hundred percent accuracy is mandatory before moving to the next phase. Second, flashcards of 8-digit codes are used to test and facilitate memory of actual coding possibilities. Again, 100% accuracy is necessary for entry into the next training phase.

Initial on-line training. Once the code has been mastered, trainees begin observations in a simplified form. Brief (3-10 second) videotape analogues are reviewed in twice-weekly meetings during the first month. On-line observations[1] are performed as well with a scan sampling method: the trainer points out target children at random from a playgroup, and each child is observed for three seconds. This scan procedure usually results in only one string of behavior being recorded per child. These observations, and other on-line coding referred to in the text, were collected from behind a one-way mirror in an adjoining observation deck.

Initial codings are done with a pencil and paper in order to allow trainees to concentrate on only one new task. A gradual progression is made whereby trainees at first code only the context, ignoring the other four portions of the code. Once item-by-item agreement reaches .95, trainees begin concentrating on the first three segments of the code. Once agreement on the three segments combined exceeds .90, the transition is made to coding the complete 5-item, 8-digit code. This procedure is

continued until item-by-item agreement is .90.

OS-3 orientation. During the weeks when initial observing is done, two 2-hour work shops are held to familiarize trainees with the OS-3, a hand-held computer designed specifically to perform general data collection and real time sampling. The internal clocks of the machines are synchronized so that direct comparisons may be made. With the assistance of an operations manual, trainees learn how to operate the OS-3 quickly. Once general operations are mastered, daily 15-minute drills are held where trainees input 8-digit codes from flashcards at increasingly rapid rates, starting with 15-second intervals and eventually working down to 2-second intervals. This drill deserves additional comment. Automated data collection devices offer the advantage of allowing observers to watch the subject continuously rather than having to attend intermittently to the recording device. We have noticed a tendency among trainees, however, to glance down at the OS-3 display screen. This in turn reduces the proportion of time they spend tracking the target. At first these glances prove useful to trainees, giving them immediate feedback on their entries, but it is a habit that could eventually lead to selective data collection and lower reliability. The flashcard exercise demands continuous vigilance, particularly as the trial length is reduced and there is less time to check entries.

Introduction of priority rules. The Interactive Code is exhaustive by design. But, like other codes, its items are not mutually exclusive; there are hierarchical components. To avoid confusion and/or the the development of individual rule sets, trainees are given priority rules that are applied when a situation involves deciding which of several items represents the "best answer." While development of priority ratings is a lengthy task, it is well worth the time invested. "Best choice" errors are far more common among our trained observers than are errors of omission or commission or highly discrepant coding. And while priority rules will not completely eliminate these miscues, they effectively minimize them.

Real time training. Following several days of scan sampling with the OS-3, trainees begin coding taped segments that require the input of two separate 8-digit entries. This allows trainees to make the qualitative shift to continuous observation with as little difficulty as possible. Within several days, trainees begin observing children in our playgroups for 15-second intervals. In this case, the average number of necessary 8-digit inputs per segment is 2.5, with a range of 1 to 5 inputs. Segment length is gradually increased over a period of several weeks, and observers are coding 3-minute intervals on-line within two weeks of when real time training started. The transition to much longer intervals (10-15 minutes) can be made within an

additional week's time in most cases.

At this point, observers begin to receive rapid feedback on agreement and reliability from a new source, via a sophisticated computer program that matches their data to that of another observer (on-line) or to a preselected criterion (videotape). Up to this point, observers have received feedback for the scan sampling and the brief analogues by hand scoring. The computer feedback is a key feature of the training program, for it provides observers with on-the-spot information as to what they are doing right and wrong. They are quickly able to determine error sources at several levels, including which portions of the code their errors are most likely to appear in, whether they are applying priority rules appropriately, and which of their errors are of omission or of commission. The program also allows us to look at group trends and potential drift, which can then be discussed in greater detail during group meetings.

Observer maintenance. Once trainees have reached our criterion for agreement (.80 x time and .80 x entry; both statistics discussed later in this paper), weekly sessions are held to check and discuss reliability, to provide individual performance feedback, and to discuss issues the observers find particularly important. Videotape and flashcard materials are intermittently used to provide immediate feedback and to emphasize particular skills or portions of the code. Although

weekly meetings can seem tedious to trainees and coordinators alike following a 12-16 week development program, the importance of these meetings should not be underestimated. Not only do these meetings serve to guarantee contact between director/coordinator and observer(s), providing an opportunity to prevent various forms of data erosion (observer drift, etc.), they allow time to deal with other problems that may arise. A colleague once characterized observational projects without weekly meetings as much like a dog sled team with no lead: everyone thinks they know what they're doing, but rarely do they end up all pulling in the same direction.

Data Collection

Following completion of the training program, observers begin collecting data on children in playgroups we run at the University in two separate playgroups: toddler groups for children 18 to 36 months of age, and infant groups for children 12 to 18 months of age. The infant groups consist of six to eight children per class, and there are typically three teachers. The toddler groups consist of 12 to 15 children per class, usually with four or five teachers. The infant play room is 3 x 5.5 meters, and the toddler room is 5 x 7.5 meters. Each room is equipped with appropriate play equipment.

The observers collect data from an adjoining room through a one-way mirror. The rooms are wired for sound, and observers

wear individual headsets to minimize distraction from any sounds originating outside the playrooms.

Assessing Observer Agreement

Continuous time data allow us to make direct assessments of observer reliability and agreement rather than correlational ones. The data from the studies reviewed by Weinrott and Jones (1984) were not real-time; only frequencies were recorded. The authors were limited in their options for testing agreement, and correlational analysis was the alternative they selected. Thus, the limitations were such that one could have, at best, made inferences about agreement based on ICCs. It would be possible, for instance, to have real-time data with neither item-by-item nor time agreement, and still report a powerful coefficient (i.e., observers could see the same things at different times). It is plausible that the reduced ICCs reported by Weinrott and Jones under covert conditions are not the results of lower levels of observer agreement, only that the patterning of inputs had somehow been altered by one or more observers. All we know from these data is that frequencies vary in a more similar fashion for pairs of observers in the overt condition, and this may be more the result of indirect cueing rather than better coding: observers undoubtedly noticed one another's entries and, therefore, could have been receiving information to the effect that their coworker felt a change had occurred. It would be

difficult to imagine this not influencing the chances of the second observer entering something new. This, of course, would lead to a more similar number of total entries than in covert assessments, and, depending on contextual variables, it could also influence the entry rate of numerous individual codes.

We use two general measures to assess observer agreement: percentage of time in agreement and percentage of entries in agreement. Both statistics are performed on each of the code's five segments (context, interactive, recipient, reactor, reaction) and for the entire 8-digit 5-item code. Data from the five subsets are useful during training for pinpointing sources of error, but they are not relevant to this discussion. Only the 8-digit agreement data will be reported. We are primarily interested in interaction processes, so this information is the most appropriate for our research. It is also a more conservative estimate of agreement than the segment data.

Time in agreement is calculated by comparing two records and dividing the number of seconds in which they have identical entries by the total number of seconds. Item-by-item agreement is checked by matching the time of a given entry in one of the two records, seven seconds before and seven seconds after, in the other record. The 7-second interval was chosen because the average entry lasts 6.2 seconds. Shorter intervals would then be meaningless in terms of matching, and longer intervals would run

into problems of chance agreement on totally different situations. If, during this time span, an identical entry is located in the second observer's record, it is considered a hit. If not, it is scored as a miss. The item-by-item assessment requires that a cross check be made, as a unidirectional comparison would allow for the possibility of extreme deflation or inflation of agreement whenever a large discrepancy in the number of entries occurred between two observers.

The rationale for adopting the time search procedure for item agreement involves several things. First, there are often differences in motor skill speed among observers, and this would lead to slightly irregular entry times. Second, and perhaps more important, the observers are frequently tracking a child in transition. It is often difficult to decide at what point a new code should be entered. This in turn can result in substantial differences among observers with respect to when an entry is made.

One argument against the time search procedure is the potential for our reliability data to be inflated by chance agreement. No doubt to some extent this is true. The kappa statistic, commonly depended upon under these circumstances, appears to be a poor predictor of interval session agreement with observational data (Mehm & Knutson, 1984). This is primarily due to variations in the level of chance agreement from one session to another. While we have not been able to determine the exact

contribution of chance to our data, we can say that no 8-digit code accounts for more than 6.0% of the total number of entries or 8.3% of the total time. This suggests that, whatever the effects of chance may be, it is unlikely that they are particularly large. It would be good, however, to determine just what the intricacies of chance are with this or any other code. Contingencies (i.e., given Code X, the probability it is preceded or followed by Code Y is ...) surely play a role in this respect. Children do not randomly partake in all the different items we record, and one might hypothesize that seasoned observers could begin to pick up on various stereotypic patterns of play that they could access and apply from time to time.

Data collection. In the laboratory we have managed to put into play a data collection system that insures covert agreement checks on a regular basis. Observers in our laboratory are each given class lists of subjects, and they are allowed to individually select which child they would like to observe at a given time, with the proviso that they not discuss their selections with other observers. Typically, there are between three and six observers coding in a room at any given time. There are several key limitations to the observer's apparent freedom in selecting. First, every observer must complete one 15-minute observation on each of the children before doing a second observation on any child. This results in an equal number

of observations for each child and helps to eliminate a source of selective bias. Second, there is a requirement that observers begin and end each 15-minute observation simultaneously. This gives us a procedure to check agreement that is not only covert but somewhat random as well. During any 15-minute interval the odds of at least one check occurring are quite high (1 - [N - 1/N x N - 2/N x N - 3/N x ...]). For instance, suppose 12 children are present in the toddler group and six observers are on hand. The probability that there would be some form of match in selection would be: 1 - (11/12 x 10/12 x 9/12 x 8/12 x 7/12) = .79. Odds in the infant room, where there are at most eight children, are even higher: 1 - (7/8 x 6/8 x 5/8 x 4/8 x 3/8) = .92.

Overt checks of agreement are also done to insure the integrity of reliability data. A matrix of observer names is posted (8 x 8), and observers are informed they should get together twice with each of their coworkers so that agreement checks on each pair of observers is guaranteed. This process is completed in about two days' time. In every case, the observers know agreement will be checked.

Results

Agreement data. Covert agreement checks, 10 from each academic quarter between July 1983 and June 1984, were randomly selected from nearly 400 such checks compiled during that

period. Covert checks came from one 8 x 8 matrix, for a total of 56 such pairs.

Table 1 includes data from covert and overt laboratory checks. Two noticable trends appear: (1) agreement ratings on the whole are quite high, and (2) there appear to be no differences in agreement based on whether assessment is made overtly or covertly in the laboratory.

Insert Table 1 about here

Discussion

As demonstrated in the previous section, it is possible to obtain observers who maintain a high degree of proficiency on a very complex coding system and to keep them at that level with a program that provides continuous monitoring of accuracy. Such a program is not without cost, but, then, highly skilled technicians are expensive in any kind of research.

Now that such procedures are available, they will undoubtedly be employed by researchers with increasing frequency. However, the availability of such systems introduces one very important problem: researchers might sometimes fail to determine if a complicated code is required to answer their research questions, yet use such a code anyway. Some questions require the use of complicated systems, and others do not. For instance, we have

recently been studying differences between the interaction patterns of abusive families versus control families in a laboratory setting. If we were merely interested in whether abusive parents give more negative and fewer positive responses than control parents, a simple frequency code would suffice. Observers for this type of study train up to 95% agreement in just six hours. However, if we wish to study the process of interaction in the abusive versus the control families, a complex code would be needed, requiring prolonged training much like that described in this paper. We would not suggest that all researchers using observation should develop complicated codes and expensive training systems, for many studies require only that differences between groups be established. If the study is designed to look at complicated interactional processes, however, a complicated code should be considered.

There is an additional reason why one should consider spending the necessary time to gather complex information. In the case of large research projects, it is difficult to predict in advance all the questions one might want to ask. This is particularly true in longitudinal projects, where the data are often analyzed by different people at different points in time. If one collects a broad spectrum of information, the very expensive collection technique will often more than pay for itself through unexpected analysis dividends. The well known

Berkeley Guidance Study (MacFarlane, 1938) serves as a case in point. This study was designed to look at how different patterns of family interaction affected children's development. The original research team decided to collect as much information on all family members as possible. Many of these data were never analyzed by the original team, which instead concentrated mainly on developmental aspects of problem behavior. Two different research teams have since used the data, however, long after the target children had grown up. Block (1971) analyzed the data from the perspective of trait theory and provided new insights into the issues of continuity and change over time. Elder, Liker, and Cross (1984) reviewed the data from yet another perspective. They described how patterns of irritability within the family interacted with the major disruption of the depression years and showed how such interactions could help explain successful and unsuccessful adaptations to periods of major change. Such additional use of the data would have been impossible had MacFarlane and her colleagues decided to use narrow-band methods of data collection.

One reason researchers to date may have been reluctant, if not unable, to use complex systems centers around the technological requirements of the procedure. Only recently has the point been reached where process data can be collected and analyzed economically. Our training procedure is expensive, but

the actual collection, storage, and analysis of the data costs far less than it did several years ago. This is due to the development of on-line data collection devices such as the OS-3 and inexpensive microcomputers that permit the immediate editing and analysis of data. A few years ago this level of complexity would have been impractical and prohibitively expensive, in terms of both time and money. Furthermore, the quality of the data using modern equipment is substantially superior. When paper and pencil techniques are used, observer errors are present just as with computerized techniques. In computerized collection, errors in transcription by keypunching have been eliminated. In our early studies we found keypunch error rates of 2% to 7%, and correcting these errors was difficult and expensive. The time line for completion of research has also sped up immensely. Where it used to take three to four months to clean up and punch observation data into the system, it is now completed in a matter of minutes from the time an observation is finished.

To summarize, changes in technology and greater attention to training and maintenance procedures produce observers who are more than capable of dealing with very complicated coding systems. Quite obviously, researchers should not choose a complicated system of observation when planning a short-term study. However, we feel that many researchers go ahead using complicated systems with relatively untrained observers. The

quality of the data then suffers, and the field loses. Surely the actual collection of data deserves as much attention as we later pay to the statistical analyses. Good observers should be hired as part of a research team, just as we hire programmers and statisticians. In fact, we would argue that, if one has to cut costs, don't do it with the actual data collection, for you can never recoup. In addition, we feel such systems should be used to facilitate studying the full complexity of interaction; otherwise, it will be difficult to make progress in understanding how social processes influence the developing child.

References

Block, J. (1971). Lives through time. Berkeley: Bancroft Books.

Elder, G. H., Liker, J. K., & Cross, C. E. (1984). Parent-child behavior in the great depression, life course and intergenerational influences. In P. B. Baltes & O. G. Brim (Eds.), Life span development and behavior (Vol. 6). New York: Academic Press.

Fagot, B. I. (1978). Reinforcing contingencies for sex role behavior: Effects of experience with children. Child Development, 49, 30-36.

Fagot, B. I., & Hagan, R. (1984). Comparisons of behavioral data from taped and on-line sequences: Is what you see what you get? Manuscript in preparation.

MacFarlane, J. W. (1938). Studies in child guidance: Methodology of data collection and organization. Monographs of the Society for Research in Child Development, 3 (6, No. 19).

Mehm, J. G., & Knutson, J. F. (1984). An empirical comparison of strategies for data analysis and assessments of observer agreement. Unpublished manuscript, University of Iowa.

Reid, J. B. (1982). Observer training in naturalistic research. In D. P. Hartmann (Ed.), Using observers to study behavior: New directions in methodology (pp. 37-50). San Francisco: Jossey Bass.

Weinrott, M. R., & Jones, R. R. (1984). Overt versus covert assessment of observer reliability. Child Development, 55, 1125-1137.

Table 1

Percent of Agreement by Item and Time for Laboratory Observations

	Percent of item agreement	Percent of time agreement
Overt check	92.1	93.6
Covert check	91.8	93.8